Review on ‘GPRChinaTemp1km: a high-resolution monthly air temperature dataset for China (1951-2020) based on machine learning’ by He et al., 2021.

The presented study aims to produce a long term, gridded product for monthly air temperature over China at high spatial resolution (1km grid spacing) based on observational data from meteorological stations operated by the China Meteorological Service. This is a challenging task since observations are scattered irregularly over the country with very limited number of station sites in Western China where spatial variability is large due too complex terrain.

Despite the appealing aims of this work which fit well in the scope of the journal, the applied methods show up with major deficiencies. The aspects are listed in the following together with some recommendations for potential improvement. Note that a couple of points have already been mentioned by the referees of the previous submission (see here).

(1) In the introduction, the advantages and disadvantages of different data sources for generating long term, gridded 2m temperature products for China are discussed. However, the inclusion of reanalysis data such as the ERA5 dataset is not considered, even though they provide consistent, dynamical information on the atmospheric state over several decades. Likewise, other relevant studies with similar aims are not mentioned as well, e.g. Peng et al., 2019 [1]

(2) The method for data splitting (random spatial split) is improper since there is strong spatial autocorrelation in the data. Especially, for flat areas with a dense observational network, this leads to an over-simplification for the interpolation task: Even with a naïve mapping to a neighboring station from the training dataset during inference (i.e. corresponding to a nearest neighbor approach), the result would be very good. In other words, the ML-model does not learn from the features and becomes incapable to generalize. Consequently, the results get much poorer, also seen from the much large residuals in regions with complex terrain and a less dense observational network (i.e. in the Himalaya region west of 100°E and south of 40°N). To avoid strong autocorrelation, data must be split along the temporal axis, i.e. by assigning sequential years of data to the test dataset, see, e.g., Section 2.2. in Kleinert et al., 2020 [2] for a more detailed discussion. Note that this aspect is also strongly related to the subsequent major points.

(3) Only static feature variables are used. The only way for the model to get dynamical information is via optimization on the predictand (the 2m observations from the training dataset). This also reasons why a model has to trained for each month of the considered period, resulting in an enormous number of models. Including (coarser) reanalysis data from a numerical atmospheric model to the predictors is therefore strongly recommended, since it allows to circumvent the spatial auto correlation problem (see aspect 2) and to generalize better. Besides, not only elevation is crucial for the near surface temperature, but also strongly the ambient topography (i.e. location in a valley or on a mountain). With the current approach, this information is missing, even though other studies point out its importance (e.g. Sha et al., 2020 [3]).
Furthermore, there are longer periods where some of the predictors are practically uncorrelated with the target quantity. Thus, it’s unlikely that these variables contribute to the interpolation model for the respective months (e.g. longitude and elevation for winter months).

(4) Evaluation of the model performance does not serve the aims of the paper, i.e. provide accurate data for regions with sparse observations. Due to the dominance of stations in the flat, densely observed parts of China, the accuracy metrics presented are too optimistic for the regions where the underlying terrain is complex and/or observations are sparse. There are only twenty stations in the Himalaya region (west of 100°E and south of 40°N), thus the results are dominated by the majority of stations located in the flat regions to the east. Alternative data splitting with an adaption of the approach (inclusion of dynamic predictors) would increase the robustness in the evaluation results for the mountainous regions.

(5) There are follow-up deficiencies in the analysis:

(a) The patterns described in Section 4.2. are large-scale patterns that does not relate to the high spatial resolutions of the dataset. The described pattern would be indeed evident in datasets with much coarser resolution, while interesting patterns due to variations in topography/ land use are not investigated.

(b) The trend analysis shows up with patterns in North-Western China for winter months that look like artefacts. The bulls-eye structure as well as the trend gradient are most likely artefacts since observational data are sparse or not existent in this region. At least, references to other work which explicitly deal with trends in the Xinjiang region would be required in addition to a more elaborated and robust method.

(c) ANUSPLIN performs better for July in the 70s, 80s and 90s (see Fig. 8). However, this is not discussed/mentioned subsequently.

(d) Discussion on the limitations does not focus on the real issues with the approach, especially regarding the static predictors.

(6) The comparison with other datasets is misleading. ERA5 and FLADS have a much coarser spatial resolution than the analyzed 1x1 km-dataset (factor of 10 and more). Thus, comparison is not straightforward and at minimum requires a reflection due to mismatches in surface elevation. Rather a comparison to other datasets with matching spatial resolution as presented in [1] would be fair and supportive.

Besides, there are couple of minor comments/issues. The most relevant are listed subsequently:

* Slitting up the data for Tmean, Tmax and Tmin into three datasets is unnecessary. Since the approach is the same, provide in a joint dataset with only one DOI.

* Dynamical and statistical downscaling techniques with reference to reanalysis datasets must be mentioned in the introduction, e.g. something like COSMO-REA2 (see,e.g., [4]) for other regions on Earth.

* Provide references to the statement on the rather poor quality of remote sensing data (see l.66)

* How is the STRM DEM data remapped onto the 1x1 km-grid? Note: should not be a bilinear, but rather an averaging method.

* Only mentioned used software once. Repeated reference to MATLAB is unnecessary. Rather spent some additional words on the technique itself.

* l.149: Should be ‘ensemble machine learning’

* l.219: Unnecessary reference to Equation 4 which directly follows the sentence.

* l.343.f.: Sentence is barely understandable and requires reformulation.

Literature reference:

[1] Peng, Shouzhang, et al. "1 km monthly temperature and precipitation dataset for China from 1901 to 2017." *Earth System Science Data* 11.4 (2019): 1931-1946. DOI.

[2] Kleinert, Felix, Martin G. Schultz, and Lukas H. Leufen. "IntelliO3-ts v1. 0: A neural network approach to predict near-surface ozone concentrations in Germany." *Geoscientific model development discussions* 2020.FZJ-2020-05012 (2020): 1-69. DOI.

[3] Sha, Yingkai, et al. "Deep-learning-based gridded downscaling of surface meteorological variables in complex terrain. Part I: Daily maximum and minimum 2-m temperature." *Journal of Applied [3] Meteorology and Climatology* 59.12 (2020): 2057-2073. DOI.

[4] Wahl, Sabrina, et al. "A novel convective-scale regional reanalysis COSMO-REA2: Improving the representation of precipitation." *Meteorol. Z* 26.4 (2017): 345-361. DOI.